Patent
Attorney's Docket No. BBNT-P01-087

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	**Mail Stop Amendment**
Jayadev Billa et al.	Group Art Unit: 2626
Application No.: 10/685,566	Examiner: J. Jackson
Filed: October 16, 2003	
For: PRONUNCIATION SYMBOLS BASED ON THE ORTHOGRAPHIC LEXICON OF A LANGUAGE	

DECLARATION UNDER 37 C.F.R. § 1.131

U.S. Patent and Trademark Office
Customer Window, Mail Stop Amendment
Randolph Building
401 Dulany Street
Alexandria, VA 22314

Sir:

I, Jayadev Billa, hereby declare that:

1. I am a co-inventor, together with Francis Kubala, of the claimed subject matter in the above-identified patent application.

2. The invention was reduced to practice prior to September 2002 in a software system called *Rough 'n 'Ready*.

3. *Rough 'n 'Ready* is a software system for the real-time indexing and browsing of broadcast news in English. The original version of the *Rough 'n 'Ready* software system was extended to include the concepts consistent with the invention.

4. The extensions to the *Rough 'n 'Ready* software system, which include concepts consistent with the invention, were described in the publication "Audio Indexing of Arabic Broadcast News," which was published at the IEEE International Conference on Acoustics,

Speech, and Signal Processing (ICASSP), Orlando, Florida, in May of 2002. The publication is attached hereto as exhibit A.

5. In particular, the publication "Audio Indexing of Arabic Broadcast News," in section 1 ("Introduction") describes the *Rough'n'Ready* system and notes that the *Rough'n'Ready* system has been extended to include the concepts described in the publication. Section 4 of the publication ("Arabic ASR") describes, among other things, aspects of the invention.

6. The *Rough'n'Ready* system, prior to September 2002, operated for its intended purpose and had a known utility.

7. Thus, the publication "Audio Indexing of Arabic Broadcast News" provides clear evidence that the invention was reduced to practice prior to September 2002.

8. I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and, further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements made jeopardize the validity of the application or any patent issuing therefrom.

Date: 7/10/2007 Signature: _____
Jayadev Billa

U.S. Patent Application No. 10/685,566
Attorney's Docket No. BBNT-P01-087

LIST OF EXHIBITS

A: Audio Indexing of Arabic Broadcast News," which was published at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, May 13-17, 2002.

Exhibit A

AUDIO INDEXING OF ARABIC BROADCAST NEWS

J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, F. Kubala

BBN Technologies
Cambridge MA 02138.
jbilla@bbn.com

ABSTRACT

This paper describes the development of the BBN Audio Indexing System for broadcast news in Arabic. Key issues addressed in this work revolve around the three major components of the audio indexing system: automatic speech recognition, speaker identification, and named entity identification. The system deals with several challenges introduced by the Arabic language, including the absence of short vowels in written text and the presence of compound words that are formed by the concatenation of certain conjunctions, prepositions, articles, and pronouns, as prefixes and suffixes to the word stem. The lack of short vowels in the transcripts prompted a novel solution that further demonstrated the power of hidden Markov models to deal with ambiguity. Another challenge was the acquisition of appropriate language modeling data, given the absence of broadcast news data for that purpose. We present performance results for all three components of the Audio Indexing System, which we believe represent the state of the art for Arabic broadcast news.

1. INTRODUCTION

During the last few years, we developed at BBN a system, called *Rough'n'Ready* [1, 2], for the real-time indexing and browsing of broadcast news in English. The system included the following technologies: speech recognition, speaker identification, named-entity extraction, topic classification, and story segmentation. The system currently runs in streaming mode, recording broadcast news data off the air, performing recognition and indexing in real-time, and storing the results in a database for later browsing over the Web.

In this paper, we report on extending the capabilities of our audio-indexing system to include the real-time audio indexing of broadcast news in Arabic. The Arabic capability currently includes the following three components: automatic speech recognition (ASR), speaker identification, and named entity extraction. Raw broadcast speech is passed first into the speech recognition component where the audio stream is analyzed for speaker changes and pauses and each speaker "turn" is converted into a word sequence. The speaker identification component then takes in the speaker-turn information, identifies known speakers by name, and clusters unknown speakers and identifies their gender. Finally, named entities representing the names of people, places, and locations, are extracted and highlighted. The end result is an enriched textual document of the spoken audio stream.

This work was supported in part by The Office of Advanced Information Technology under Contract 1999-S0189000-000 and by DARPA under Contract N66001-00-C-8008.

First, we provide a brief primer on the Arabic language and the challenges it represents for speech recognition, followed by an overview of the the Arabic resources used in the development of the Audio Indexer components. In order, we then detail the development of each system component, starting with the Arabic ASR component, moving to speaker identification, and ending with the named entity extraction component.

2. ARABIC LANGUAGE

The various colloquial dialects that are spoken across the Arab World are very different from each other in pronunciation, vocabulary, and grammar; they are typically not written at all; and they all differ markedly from the only written language, known as Modern Standard Arabic (MSA). In addition to being the only written language, MSA is also used as the dominant language for broadcast news and for formal speeches. So, except for some pronunciation differences, the MSA spoken in Saudi Arabia, for example, is the same as that spoken in Syria, Egypt, and Morocco. Below, we use the term Arabic to refer to MSA only.

Arabic has a system of 25 consonants, three quasi-consonants (w, y, and glottal stop or *hamza*), and three vowels (a, u, i). Duration is phonemic in Arabic (i.e., a difference in the duration of a vowel or a consonant can change the meaning of a word), so vowels can be either short or long, and the consonants can be doubled. Arabic does not have regular letters to represent short vowels and the doubling of consonants; these, when written, appear as marks, or diacritics, above or below the consonants. The long vowels are always written, but the symbols used for the long vowels are also used to write /w/, /y/, and the hamza, so that creates an ambiguity between what is written and how those letters are pronounced. (The hamza takes on several forms, one of which is placed above or below the *alif*, the character used to represent the long vowel /aa/.) In ordinary Arabic text, like newspapers and most books, diacritics, including the short vowels and the doubling of consonants, are almost never written, except sometimes to disambiguate a word. Morphologically, Arabic words consist of CV and CVC syllables and the language has a basic verb-subject-object structure.

As an example of the ambiguities that are inherent in the writing system, the word that appears as "ktb" can be either "katab" (he wrote), "kattab" (he dictated), "kutib" (it was written), "kuttib" (it was dictated), or "kutub" (books). Which of the words is meant has to be inferred by the reader from the context.

Arabic is a highly inflected language, with gender, number, person, tense, mood, and case, are all indicated as prefixes, infixes, or suffixes to a basic root. The resulting *word stem* can then be "enlarged" by other prefixes and suffixes. All single-letter con-

junctions and prepositions that precede a word stem, as well as the definite article, are attached to the word stem as prefixes, and any following pronoun is attached to the word as a suffix. So, what is considered as a single word in Arabic can actually represent a whole phrase. For example, the phrase "and he will write it" is written as a single word in Arabic "wsyktbh": "w" = and, "s" = will, "y" is a marker for 3rd person singular masculine, "ktb" = write, and "h" = it.

When reading, an Arabic reader has the freedom to pronounce or delete any vowel case marker at the end of a word. There are cross-word assimilation phenomena which are also optional. These effects complicate correspondences between the written form and what is spoken.

The ambiguities mentioned above, as well as the optional pronunciations exercised by each speaker, introduce difficulties for Arabic ASR, both for acoustic modeling and training, as well as language modeling.

3. ARABIC DATA RESOURCES

Acoustic data for the ASR models consisted of approximately 42 hours of spoken Arabic, transcribed without diacritic markings, from Egyptian and Syrian broadcast radio and TV. The 42 hours of acoustic data were split into a training set and a test set. The test set of approximately 4.5 hours was chosen to be temporally disjoint from and later than training. Towards the end of this work, an additional 20 hours of broadcast news, from the Al-Jazeera TV network, was added to the acoustic training.

Language-modeling data was acquired from two Lebanese newspapers, Al-Hayat and An-Nahar, and one newswire source, the Arabic service of Agence France Presse (AFP) newswire service. Al-Hayat data consisted of four years of newspaper text spanning 1997-2000, whereas the An-Nahar data consisted of four years spanning 1996-1999. The AFP newswire data consisted of seven years of newswire text spanning 1994-2000 and was obtained from the Linguistic Data Consortium (LDC).

Table 1 shows the sizes of the acoustic training and test sets and the three-language modeling sources as measured by total number of word tokens and number of unique words.

Source	Total Words	Unique Words
Al Hayat	75M	900k
An-Nahar	74M	850k
AFP	65M	550k
Acoustic Training	330k	38k
All Training	215M	1.3M
Test	25k	8k

Table 1. Arabic Data Resources.

4. ARABIC ASR

4.1. Initial ASR System

Our ASR system is engineered for real-time and derived from the BBN Byblos system [3]. In this system, each phoneme is modeled by a 5-state HMM. But, before we could start developing the Arabic system, we had to make a number of modeling decisions that were affected by the peculiarities of the language and the data.

First of all, since the transcriptions of the audio data did not include the short vowels and consonant doubling, and did not resolve the various pronunciation ambiguities mentioned above, we had to decide whether to produce accurate phonetic transcriptions that corresponded to the audio or not. We decided *not* to produce accurate phonetic transcriptions and to use the given transcriptions as is. One reason for this decision was that all the language modeling data from newspaper text was not diacritized either, so we thought it would be easier to be consistent throughout. (Diacritizing the language modeling data would have been extraordinarily costly and impractical. Furthermore, commercially available diacritizing software was not accurate enough to be usable for our purposes.) Another reason was that, even if accurate phonetic transcriptions were available, speaker optionality in pronouncing word endings would make it difficult to provide a consistent phonetic dictionary.

So, we simply assigned one phoneme model to each character. In this manner, a model for a consonant would actually be modeling that consonant alone, as well as that consonant followed by each of the three vowels, or whatever actually occurred in the data. This assignment would also work well for word endings where the vowels were optional anyway. No allowance was made for pronunciation differences across different countries. Although Arabic is a particularly good language for making such modeling assignments because of the phonetic regularity of its writing system, we believe it could work well for other languages as well.

Another major decision had to do with compound words – words with multiple affixes attached to the stem. Because many of those words are essentially phrases, treating each of those words as a unique word results in a very strong language model (we use a statistical trigram model), thus improving recognition accuracy. However, simultaneously, we were concerned that the resulting out-of-vocabulary (OOV) word rate would be too high. With a 65,000-word lexicon, we measured the OOV rate in the test set at 4%, which is certainly on the high side. In comparison, for English broadcast news, the same size lexicon results in an OOV rate of less than 1%. One choice we had was to perform automatic stemming of the compound words to separate the various affixes from the word stem. Unfortunately, we were unable to find automatic stemming software that was accurate enough for our purposes. Therefore, we decided on keeping each compound word as a single word and eventually increased our vocabulary size to reduce the OOV rate. For example, with a 128,000-word vocabulary, the OOV rate went down to 2.4%.

Our first ASR system, based on this approach of considering each Arabic character as a phoneme, ignoring the need to stem, and allowing the acoustic models to model short vowels implicitly, performed surprisingly well. This initial ASR system, running sub-real-time, using a 65,000-word lexicon and the acoustic transcripts of the acoustic training data as the sole language modeling data (only 230k words without the Al-Jazeera data), gave a baseline word error rate (WER) of 31.2%, which was very encouraging. So, we did additional work to improve performance.

4.2. Language Modeling

Following our initial results, we focused on potential improvements with better language modeling using language data obtained from Al-Hayat and AFP. First, the text was stripped of all formatting information and punctuation. In Arabic, digits can be spoken in a number of ways depending on the kind of number (date, monetary amount, quantity, etc.) and context. The multiplicity of

possible pronunciations preempted any rational attempt to convert digits to words. We therefore mapped all digits to an unknown-word marker. In all, approximately 145 million words were added to the initial 290k words of language modeling data and the models were retrained. There was a substantial improvement in system performance with WER dropping from 31.2% to 26.6%. Results are summarized in Table 2.

4.3. System Improvements

To build on the initial results, we first attempted to improve the core acoustic models with a combination of better description of Arabic acoustic categories for clustering, and larger, more complex acoustic models.

To further improve performance, two new capabilities were added to the BBN Audio Indexer ASR component: Maximum Likelihood Linear Regression (MLLR) and the ability to use lexicons greater than 65,000 words.

MLLR [4] is a common technique used to adapt speech recognition system models to an incoming speaker. However, MLLR has hitherto been used chiefly in research systems where system speed was a tangential distraction. BBN has now implemented MLLR within the core speech recognition system with essentially no penalty in speed. The current system still runs faster than real-time with up to a 10% reduction in word error rate.

To consolidate our improvements we tuned and optimized our real-time system for best possible performance. We found a very large improvement (10% reduction in WER) for tuning and optimizing the ASR component, indicating a sub-optimal starting point. The result of all the above improvements was a reduction in the WER from 26.6% to 21.0%.

We then augmented our system with the capability to run with lexicon sizes greater than 65,000 words, again in faster than real-time. We created a 128,000 word lexicon based on the most frequent words in the three language modeling datasets (Al-Hayat, An-Nahar and AFP) and all the words in our acoustic training set. The larger lexicon contributed to a further reduction in the WER to 20.4%.

Later in our work, we received about twenty hours of transcribed broadcast news from the Al-Jazeera TV network which were added to the acoustic training corpus. This additional data provided another incremental improvement in performance, resulting in a final and very satisfying WER of 19.1%, as shown in Table 2, especially for a system that runs in real-time. The system is currently operational at BBN and receiving streaming audio from Al-Jazeera Arabic broadcast news.

System	WER
Baseline	31.2
+ 145M word LM	26.6
+ System Improvements	21.0
+ 128k Lexicon	20.4
+ Additional data	19.1

Table 2. Improvements to ASR component performance with refinements. Each row indicates incremental improvements over the component in the preceding row.

4.4. Scoring Issues

One of the liberties that writers of Arabic allow themselves is in the placement of the hamza when it occurs above or below the alif character. In all, there are three variations of hamza placement, plus the alif alone. Writers often strip the hamza and leave the alif alone, but they are not consistent in doing that. Since each of the four forms is treated as a separate character in our system, the resulting lexicon and the transcribed data are not always consistent, therefore resulting in recognition errors that generally are of little consequence to the Arabic reader.

So, we performed an experiment where the reference and hypothesized word sequences were changed to remove orthographic variations due to hamza/alif forms. The resulting WER of 17.1% in the second row of Table 2 is to be compared with the previous error rate of 19.1%, showing the prevalence of hamza/alif inconsistencies in Arabic writing. We then performed another experiment where we changed the actual training data by removing all hamza/alif orthographic variations and retrained the ASR system. The WER was reduced further to 16.5%.

System	WER
Most conservative system	19.1
Change reference and hypothesis only	17.1
Change transcripts and retrain ASR system	16.5

Table 3. Effect of changes in scoring methodology for Arabic ASR.

5. SPEAKER IDENTIFICATION

The Speaker Identification component in the Arabic Audio Indexer was modeled on the English version of the system [5, 6]. For our initial experiment, a total of eighteen speakers, eight female and ten male, were selected by picking the speakers with the most training data in the acoustic data. For consistency, the training and test set was kept identical to the one used in the ASR component.

On evaluation, the speaker identification component was able to identify all segments correctly without any false accepts or false rejects.

This result, while satisfying, has its caveats; the test set was unrealistically small with high quality audio and we had no cross-validation test so the speaker identification component was optimized on the same data it was tested on. It is prudent to expect the system to degrade when presented with more speakers and poorer acoustics.

To test this idea further, a total of two hundred and twenty nine speakers (189 male, 41 female) were selected, by picking the speakers that had at least three minutes of training data. The selected speakers were distributed over Cairo Radio, Cairo TV, Arabic VOA, Damascus Radio and Al-Jazeera TV network. Again, for consistency, the training and test sets were kept identical to the ones used in the ASR component.

On evaluation, the speaker identification component gave a segment error rate of 12.2%, i.e., the speaker was identified correctly in 87.8% of all test segments. This is an unoptimized result and should improve with tuning.

6. NAMED ENTITY EXTRACTION

The Name Entity Extraction component was also modeled on the English version of the Audio Indexer [7]. However, since various assumptions the tools make for the English version were no longer applicable, we had to modify the system for use on Arabic. The training and test sets were again kept identical to the ones used for ASR component development.

The training set consisted of approximately 3700 named entities and about 800 named entities in test. In Arabic, named entities are often combined with prefixes, typically "w" (and), "b" (with) and "l" (for), and are normally written as single words. Since this is quite different from English, the transcribers were asked to separate the prefix from each named entity. We found that, for the most part, the prefixes were accurately separated from the named entity, but, inconsistencies remain. The final breakdown of named entities is given in Table 4.

Named Entity Type	Training	Test
Persons	1899	307
Locations	1224	378
Organizations	630	118
Total	3753	803

Table 4. Distribution of named entities by type over training and test set.

The named-entity extraction component gave an F-measure (harmonic mean of precision and recall) of 90.26, measured on the original text, i.e., without any speech recognition errors. This translates to a slot-error-rate (SER) [8] of 17.0%. These results are quite impressive when one considers the fact that both the F-measure and slot-error-rate numbers are close to the best-reported results on English broadcast news text.

To further investigate the effect of the inconsistent prefix separation, we reran the system on data where we manually re-attached the separated prefix to each named entity and also where we manually and consistently separated the three most common prefixes (w, b and l) from each named entity. Table 5 summarizes the results of the three experiments. The first experiment is the baseline system with all the inconsistencies in prefix separation kept as-is during training and testing. In the second experiment, we manually re-attached all separated prefixes to its corresponding named entity, and retrained and tested the system. This increased the errors somewhat to an F-measure of 89.59 or SER of 18.0%. The last experiment was the reverse of the second experiment in that we manually separated the three common prefixes (w,b and l) in the transcripts to ensure consistent prefix separation and repeated the experiment. This final experiment gave the best results, with an F-measure of 91.25 and SER of 15.2%. The results are expected since each compound name (prefix + named entity) results in a "new" named entity within our system causing both an increase in named entities as well as a reduction in training data per named entity. However, in real life, such separation of prefix from named entity is highly improbable and the results of the second experiment, i.e., with no prefix separation, is closest to what one might expect in normal usage.

System	F-measure	%SER
Baseline w/ inconsistent prefix separation	90.26	17.0
with no prefix separation	89.59	18.0
with consistent prefix separation	91.25	15.2

Table 5. Named Entity component performance with variations on the treatment of named entity prefixes in Arabic.

7. CONCLUSIONS

This work focused on the rapid (four month) development of an Audio Indexing system for use on Arabic broadcast news. Major challenges introduced by the Arabic language, such as the absence of short vowels in written text and the presence of compound words were addressed. Our solutions to these language issues yielded what we believe to be a state-of-the-art Audio Indexing System for Arabic Broadcast news. This system currently runs in streaming mode, recording broadcast news data off the air, performing recognition and indexing in real-time, and storing the results in a database for later browsing over the Web.

8. REFERENCES

[1] Francis Kubala, Sean Colbath, Daben Liu, Amit Srivastava, and John Makhoul, "Integrated technologies for indexing spoken language," in *Communications of the ACM*. February 2000, vol. 43, pp. 48–56, ACM.

[2] John Makhoul, Francis Kubala, Timothy Leek, Daben Liu, Long Nguyen, Richard Schwartz, and Amit Srivastava, "Speech and language technologies for audio indexing," in *Proceedings of the IEEE*. August 2000, vol. 88, pp. 1338–1353, IEEE.

[3] L. Nguyen, S. Matsoukas, J. Davenport, J. Billa, R. Schwartz, and J. Makhoul, "The 1999 BBN Byblos 10xRT broadcast news transcription system," in *Proceedings of the 2000 NIST Speech Transcription Workshop*, Maryland, May 2000, NIST.

[4] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," Technical Report CUED/F-INFENG/TR. 291, Cambridge University, Engineering Department, Cambridge, England, 1997.

[5] Herb Gish and Michael Schmidt, "Text-independent speaker identification," in *IEEE Signal Processing Magazine*. October 1994, pp. 18–32, IEEE.

[6] Daben Liu and Francis Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proceedings of Eurospeech99*, Budapest, September 1999, pp. 1031–1034.

[7] D. M. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," in *Machine Learning*, 1999, vol. 34, pp. 211–231.

[8] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel, "Performance measures for information extraction," in *Proceedings of the DARPA Workshop on Broadcast News Understanding*. March 1999, pp. 37–40, Morgan Kaufmann.